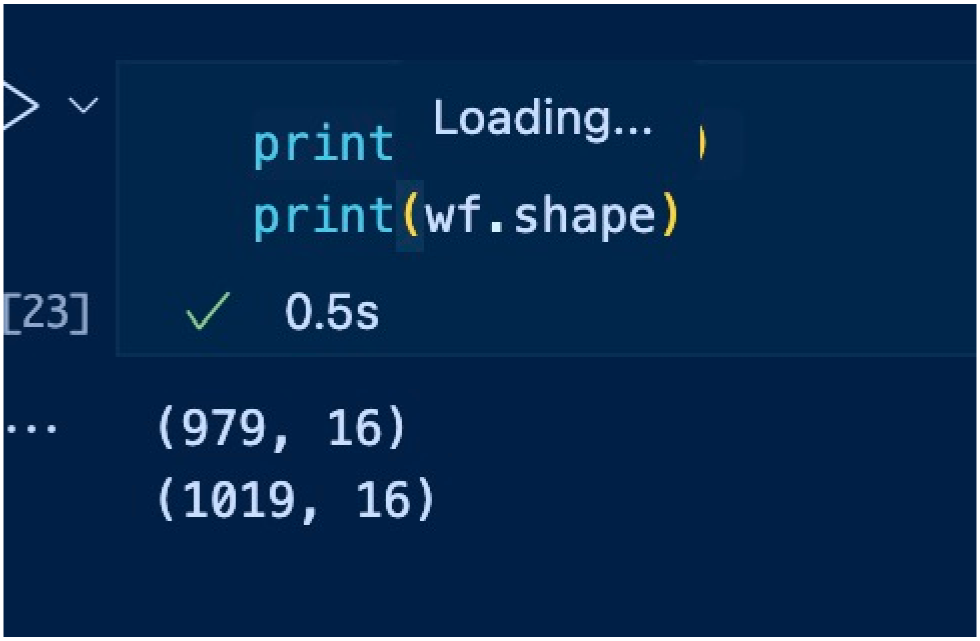
Cleaning Hubspot

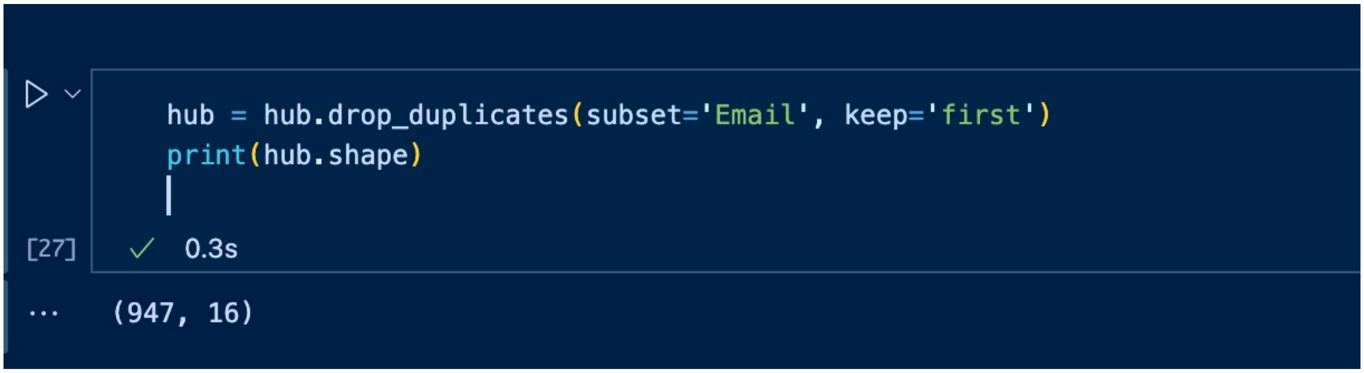
Before removing duplicate emails 979 records in Hubspot:



```
print Loading... )
print(wf.shape)
```

```
(979, 16)
(1019, 16)
```

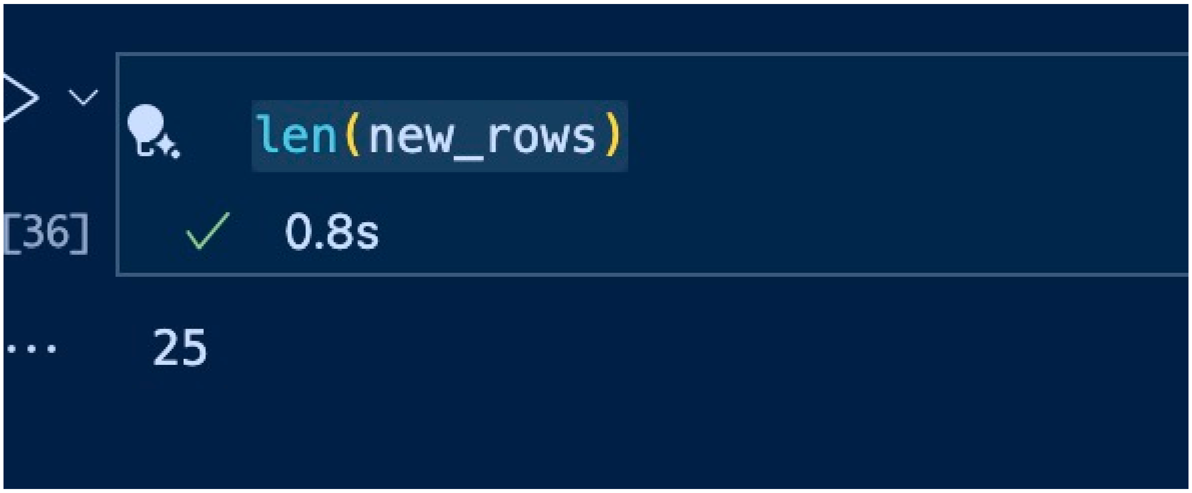
After removing duplicate emails we get 947



```
hub = hub.drop_duplicates(subset='Email', keep='first')
print(hub.shape)
```

```
(947, 16)
```

```
# Getting all rows from webflow that are not in hubspot
new_rows = wf[~wf['Email'].isin(hub['Email'])]
```

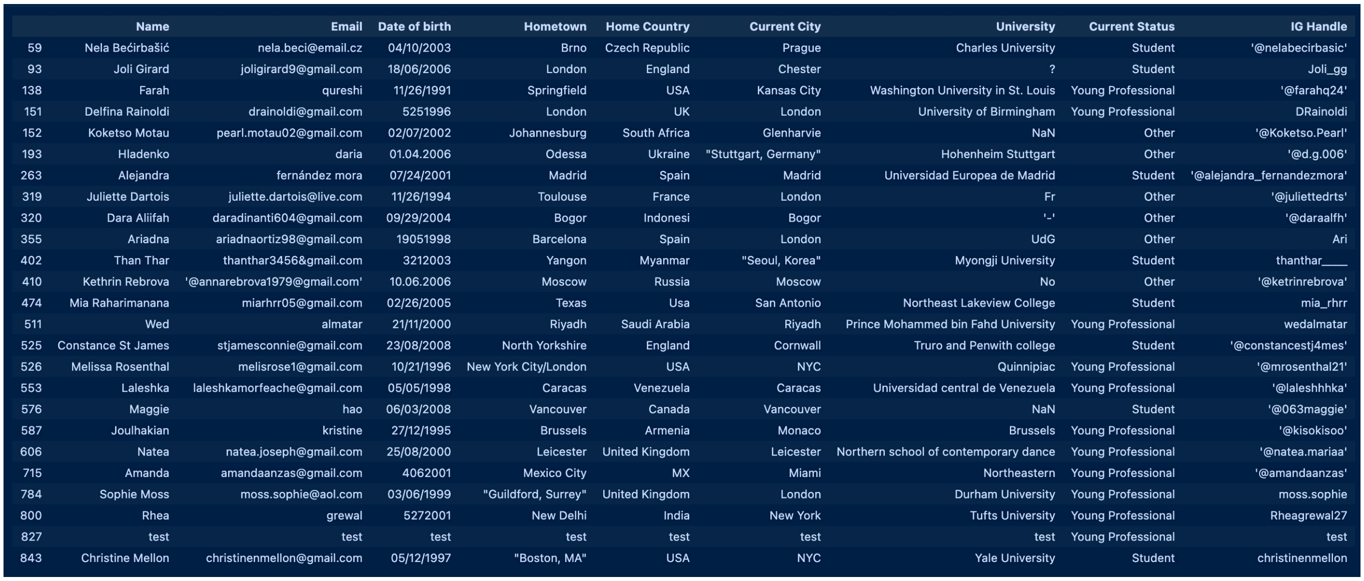


```
len(new_rows)
[36]  ✓ 0.8s
...   25
```

Adding missing rows from WF

25 rows total

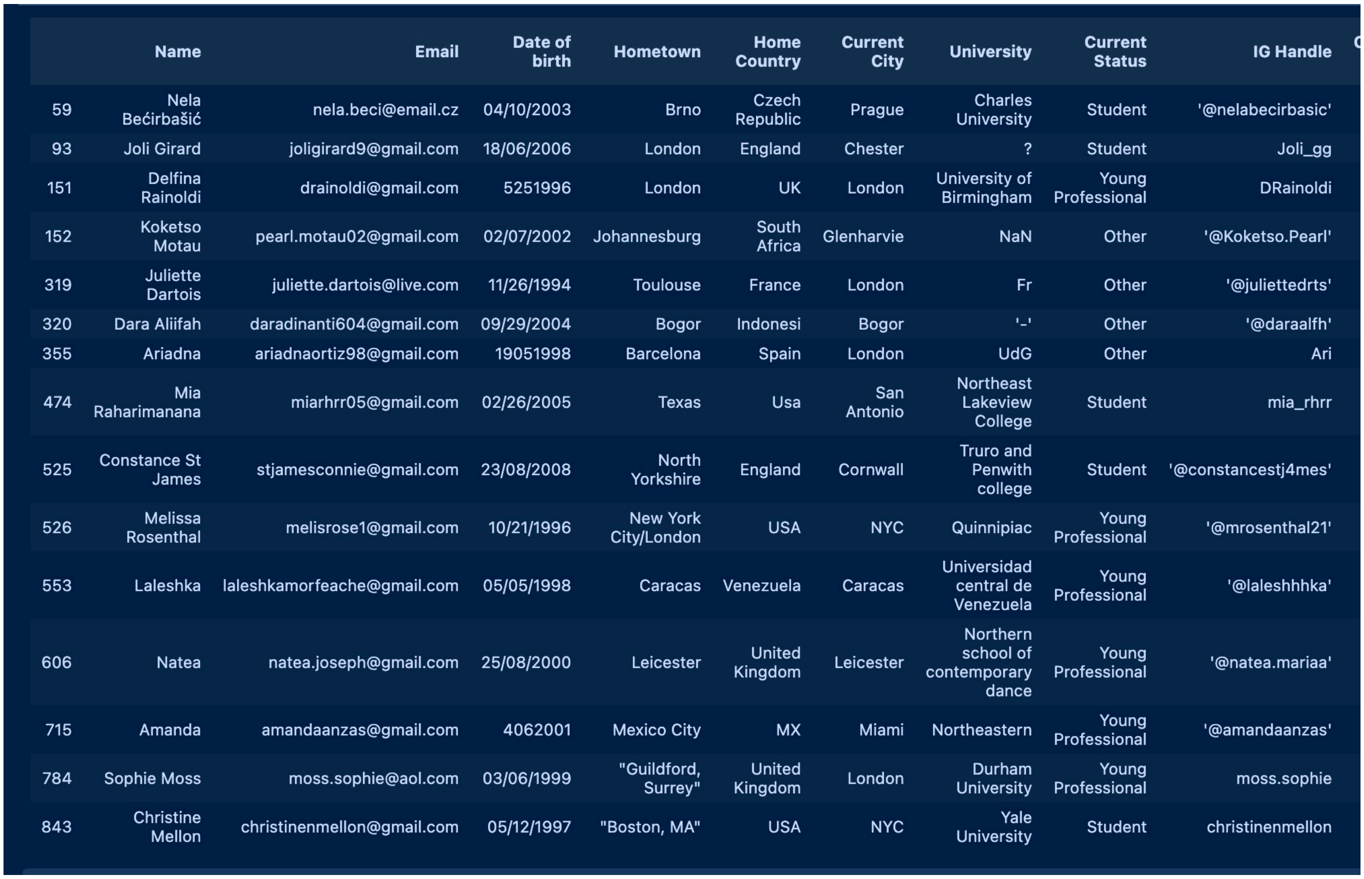
Some emails might not be valid, must check for those. For example "daria" is not a valid email:



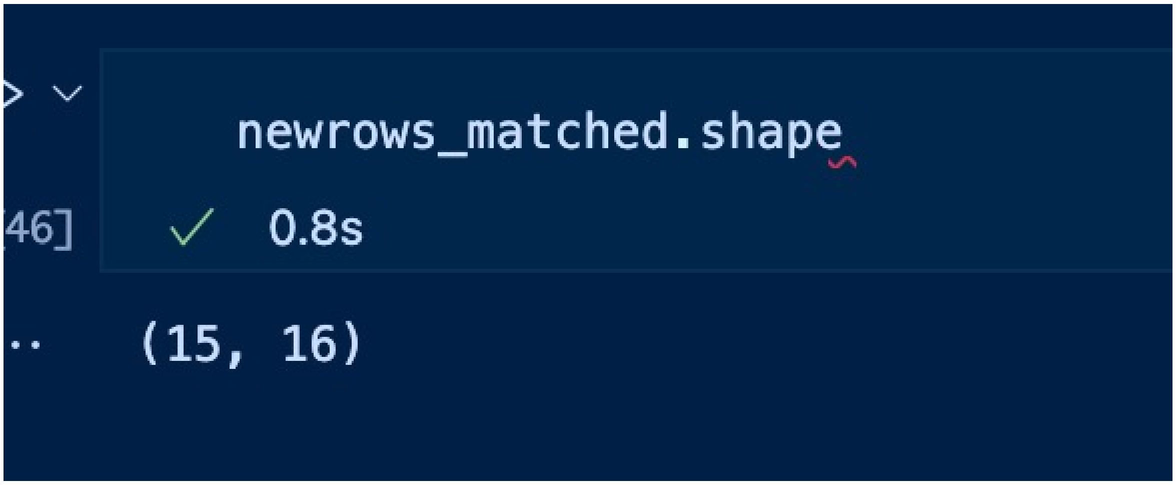
	Name	Email	Date of birth	Hometown	Home Country	Current City	University	Current Status	IG Handle
59	Nela Bećirbašić	nela.beci@email.cz	04/10/2003	Brno	Czech Republic	Prague	Charles University	Student	'@nelabecirbasic'
93	Joli Girard	joligirard9@gmail.com	18/06/2006	London	England	Chester	?	Student	Joli_gg
138	Farah	qureshi	11/26/1991	Springfield	USA	Kansas City	Washington University in St. Louis	Young Professional	'@farahq24'
151	Delfina Rainoldi	drainoldi@gmail.com	5251996	London	UK	London	University of Birmingham	Young Professional	DRainoldi
152	Koketso Motau	pearl.motau02@gmail.com	02/07/2002	Johannesburg	South Africa	Glenharvie	NaN	Other	'@Koketso.Pearl'
193	Hladenko	daria	01.04.2006	Odessa	Ukraine	"Stuttgart, Germany"	Hohenheim Stuttgart	Other	'@d.g.006'
263	Alejandra	fernández mora	07/24/2001	Madrid	Spain	Madrid	Universidad Europea de Madrid	Student	'@alejandra_fernandezmora'
319	Juliette Dartois	juliette.dartois@live.com	11/26/1994	Toulouse	France	London	Fr	Other	'@juliettedrts'
320	Dara Aliifah	daradinanti604@gmail.com	09/29/2004	Bogor	Indonesi	Bogor	'-'	Other	'@daraalfh'
355	Ariadna	ariadnaortiz98@gmail.com	19051998	Barcelona	Spain	London	UdG	Other	Ari
402	Than Thar	thanthar3456&gmail.com	3212003	Yangon	Myanmar	"Seoul, Korea"	Myongji University	Student	thanthar____
410	Kethrin Rebrova	'@annarebrova1979@gmail.com'	10.06.2006	Moscow	Russia	Moscow	No	Other	'@ketrinrebrova'
474	Mia Raharimanana	miarhrr05@gmail.com	02/26/2005	Texas	Usa	San Antonio	Northeast Lakeview College	Student	mia_rhrr
511	Wed	almatar	21/11/2000	Riyadh	Saudi Arabia	Riyadh	Prince Mohammed bin Fahd University	Young Professional	wedalmatar
525	Constance St James	stjamesconnie@gmail.com	23/08/2008	North Yorkshire	England	Cornwall	Truro and Penwith college	Student	'@constancestj4mes'
526	Melissa Rosenthal	melisrose1@gmail.com	10/21/1996	New York City/London	USA	NYC	Quinnipiac	Young Professional	'@mrosenthal21'
553	Laleshka	laleshkamorfeache@gmail.com	05/05/1998	Caracas	Venezuela	Caracas	Universidad central de Venezuela	Young Professional	'@laleshhhka'
576	Maggie	hao	06/03/2008	Vancouver	Canada	Vancouver	NaN	Student	'@063maggie'
587	Joulhakian	kristine	27/12/1995	Brussels	Armenia	Monaco	Brussels	Young Professional	'@kisokisoo'
606	Natea	natea.joseph@gmail.com	25/08/2000	Leicester	United Kingdom	Leicester	Northern school of contemporary dance	Young Professional	'@natea.mariaa'
715	Amanda	amandaanzas@gmail.com	4062001	Mexico City	MX	Miami	Northeastern	Young Professional	'@amandaanzas'
784	Sophie Moss	moss.sophie@aol.com	03/06/1999	"Guildford, Surrey"	United Kingdom	London	Durham University	Young Professional	moss.sophie
800	Rhea	grewal	5272001	New Delhi	India	New York	Tufts University	Young Professional	Rheagrewal27
827	test	test	test	test	test	test	test	Young Professional	test
843	Christine Mellon	christinenmellon@gmail.com	05/12/1997	"Boston, MA"	USA	NYC	Yale University	Student	christinenmellon

```
import re

# Define a function to check if an email is valid
def is_valid_email(email):
    if pd.isna(email):
        return False
    email_regex = r'^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$'
    return re.match(email_regex, email) is not None

# Apply the function to filter out rows with invalid emails
newrows_matched =
newrows_matched[newrows_matched['Email'].apply(is_valid_email)]
```

result gives us 15 unmatched rows from webflow that were not in our original hubspot

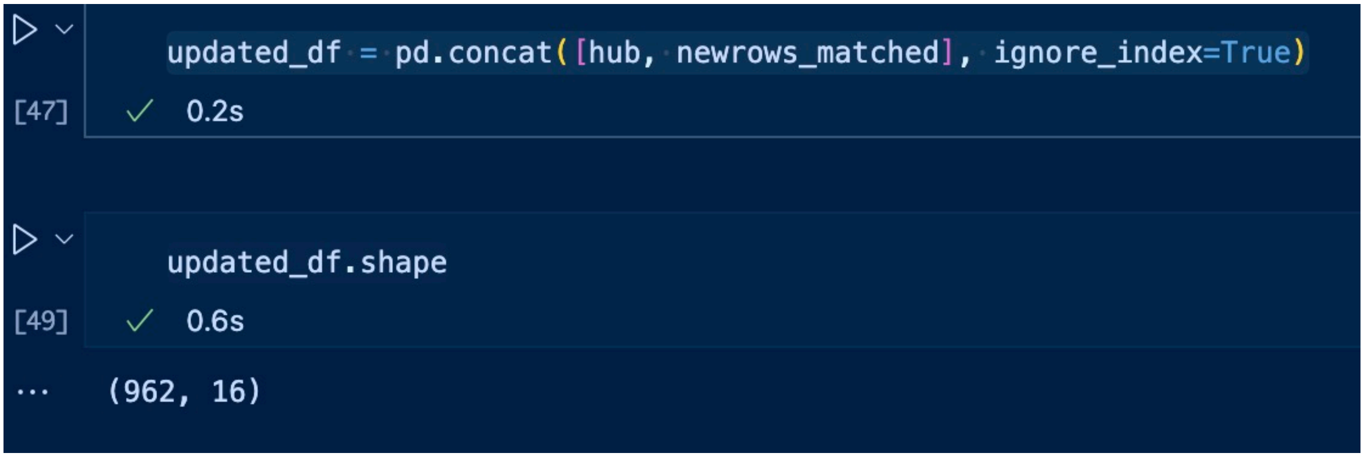


	Name	Email	Date of birth	Hometown	Home Country	Current City	University	Current Status	IG Handle
59	Nela Bećirbašić	nela.beci@email.cz	04/10/2003	Brno	Czech Republic	Prague	Charles University	Student	'@nelabecirbasic'
93	Joli Girard	joligirard9@gmail.com	18/06/2006	London	England	Chester	?	Student	Joli_gg
151	Delfina Rainoldi	drainoldi@gmail.com	5251996	London	UK	London	University of Birmingham	Young Professional	DRainoldi
152	Koketso Motau	pearl.motau02@gmail.com	02/07/2002	Johannesburg	South Africa	Glenharvie	NaN	Other	'@Koketso.Pearl'
319	Juliette Dartois	juliette.dartois@live.com	11/26/1994	Toulouse	France	London	Fr	Other	'@juliettedrts'
320	Dara Aliifah	daradinanti604@gmail.com	09/29/2004	Bogor	Indonesi	Bogor	'-'	Other	'@daraalfh'
355	Ariadna	ariadnaortiz98@gmail.com	19051998	Barcelona	Spain	London	UdG	Other	Ari
474	Mia Raharimanana	miarhrr05@gmail.com	02/26/2005	Texas	Usa	San Antonio	Northeast Lakeview College	Student	mia_rhrr
525	Constance St James	stjamesconnie@gmail.com	23/08/2008	North Yorkshire	England	Cornwall	Truro and Penwith college	Student	'@constancestj4mes'
526	Melissa Rosenthal	melisrose1@gmail.com	10/21/1996	New York City/London	USA	NYC	Quinnipiac	Young Professional	'@mrosenthal21'
553	Laleshka	laleshkamorfeache@gmail.com	05/05/1998	Caracas	Venezuela	Caracas	Universidad central de Venezuela	Young Professional	'@laleshhhka'
606	Natea	natea.joseph@gmail.com	25/08/2000	Leicester	United Kingdom	Leicester	Northern school of contemporary dance	Young Professional	'@natea.mariaa'
715	Amanda	amandaanzas@gmail.com	4062001	Mexico City	MX	Miami	Northeastern	Young Professional	'@amandaanzas'
784	Sophie Moss	moss.sophie@aol.com	03/06/1999	"Guildford, Surrey"	United Kingdom	London	Durham University	Young Professional	moss.sophie
843	Christine Mellon	christinenmellon@gmail.com	05/12/1997	"Boston, MA"	USA	NYC	Yale University	Student	christinenmellon



```
newrows_matched.shape
[46]  ✓  0.8s

(15, 16)
```

Adding to the hubspot gives 962 total rows:



```
updated_df = pd.concat([hub, newrows_matched], ignore_index=True)
```

```
[47] ✓ 0.2s
```

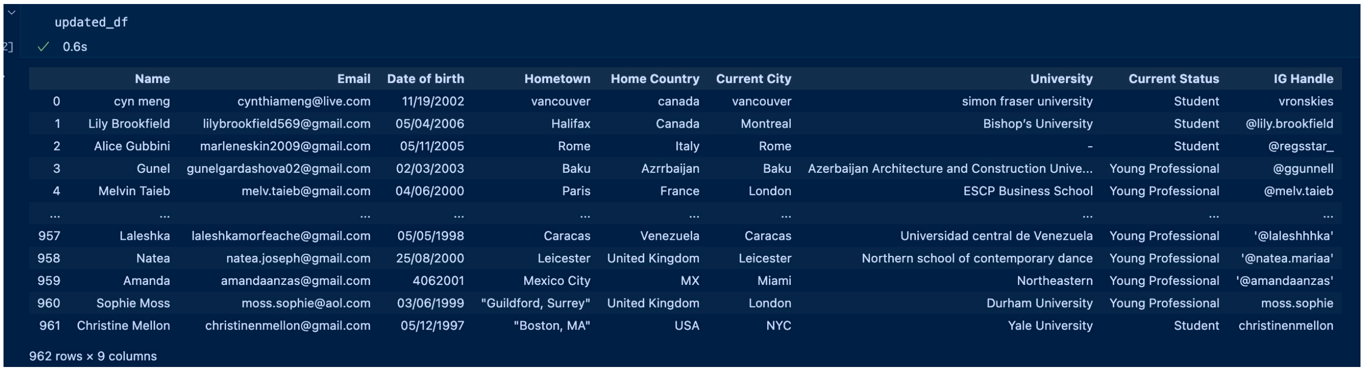
```
updated_df.shape
```

```
[49] ✓ 0.6s

... (962, 16)
```

Recleaning hubspot

We might not need every column, so we are only keeping a subset:

```
updated_df = updated_df.drop(columns=['Conversion Date', 'Conversion Page', 'Conversion Title', 'Contact first name', 'Contact last name', 'Contact email', 'Contact ID'])
```

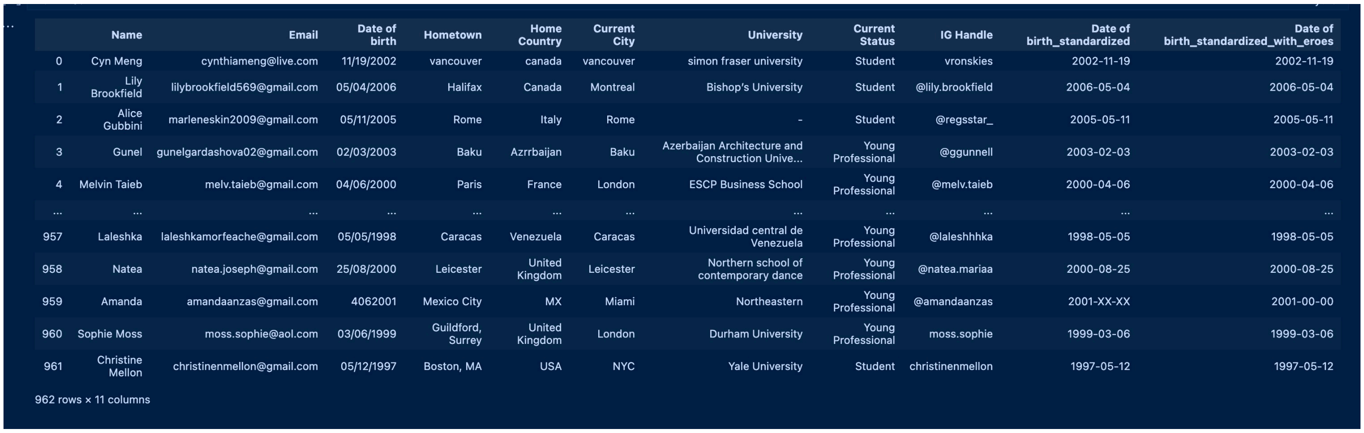


```
updated_df
✓ 0.6s
```

	Name	Email	Date of birth	Hometown	Home Country	Current City	University	Current Status	IG Handle
0	cyn meng	cynthiameng@live.com	11/19/2002	vancouver	canada	vancouver	simon fraser university	Student	vronskies
1	Lily Brookfield	lilybrookfield569@gmail.com	05/04/2006	Halifax	Canada	Montreal	Bishop’s University	Student	@lily.brookfield
2	Alice Gubbini	marleneskin2009@gmail.com	05/11/2005	Rome	Italy	Rome	-	Student	@regsstar_
3	Gunel	gunelgardashova02@gmail.com	02/03/2003	Baku	Azrrbaijan	Baku	Azerbaijan Architecture and Construction Unive...	Young Professional	@ggunnell
4	Melvin Taieb	melv.taieb@gmail.com	04/06/2000	Paris	France	London	ESCP Business School	Young Professional	@melv.taieb
...	...	...	...	...	...	...	...	...	...
957	Laleshka	laleshkamorfeache@gmail.com	05/05/1998	Caracas	Venezuela	Caracas	Universidad central de Venezuela	Young Professional	'@laleshhhka'
958	Natea	natea.joseph@gmail.com	25/08/2000	Leicester	United Kingdom	Leicester	Northern school of contemporary dance	Young Professional	'@natea.mariaa'
959	Amanda	amandaanzas@gmail.com	4062001	Mexico City	MX	Miami	Northeastern	Young Professional	'@amandaanzas'
960	Sophie Moss	moss.sophie@aol.com	03/06/1999	"Guildford, Surrey"	United Kingdom	London	Durham University	Young Professional	moss.sophie
961	Christine Mellon	christinenmellon@gmail.com	05/12/1997	"Boston, MA"	USA	NYC	Yale University	Student	christinenmellon

962 rows × 9 columns

Removing more inconsistencies like quotes around cities and @ in IG handle.
Some dates are completely wrong so applying best guess to normalize.



	Name	Email	Date of birth	Hometown	Home Country	Current City	University	Current Status	IG Handle	Date of birth_standardized	Date of birth_standardized_with_eroes
0	Cyn Meng	cynthiameng@live.com	11/19/2002	vancouver	canada	vancouver	simon fraser university	Student	vronskies	2002-11-19	2002-11-19
1	Lily Brookfield	lilybrookfield569@gmail.com	05/04/2006	Halifax	Canada	Montreal	Bishop’s University	Student	@lily.brookfield	2006-05-04	2006-05-04
2	Alice Gubbini	marleneskin2009@gmail.com	05/11/2005	Rome	Italy	Rome	-	Student	@regsstar_	2005-05-11	2005-05-11
3	Gunel	gunelgardashova02@gmail.com	02/03/2003	Baku	Azrrbaijan	Baku	Azerbaijan Architecture and Construction Unive...	Young Professional	@ggunnell	2003-02-03	2003-02-03
4	Melvin Taieb	melv.taieb@gmail.com	04/06/2000	Paris	France	London	ESCP Business School	Young Professional	@melv.taieb	2000-04-06	2000-04-06
...	...	...	...	...	...	...	...	...	...	...	...
957	Laleshka	laleshkamorfeache@gmail.com	05/05/1998	Caracas	Venezuela	Caracas	Universidad central de Venezuela	Young Professional	@laleshhhka	1998-05-05	1998-05-05
958	Natea	natea.joseph@gmail.com	25/08/2000	Leicester	United Kingdom	Leicester	Northern school of contemporary dance	Young Professional	@natea.mariaa	2000-08-25	2000-08-25
959	Amanda	amandaanzas@gmail.com	4062001	Mexico City	MX	Miami	Northeastern	Young Professional	@amandaanzas	2001-XX-XX	2001-00-00
960	Sophie Moss	moss.sophie@aol.com	03/06/1999	Guildford, Surrey	United Kingdom	London	Durham University	Young Professional	moss.sophie	1999-03-06	1999-03-06
961	Christine Mellon	christinenmellon@gmail.com	05/12/1997	Boston, MA	USA	NYC	Yale University	Student	christinenmellon	1997-05-12	1997-05-12

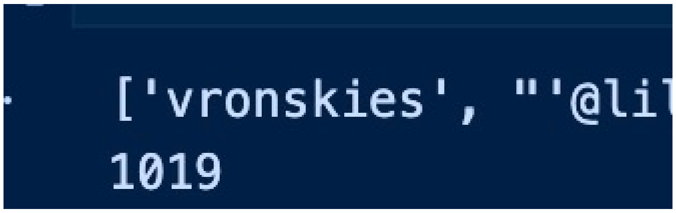
962 rows × 11 columns

IG Handles

Getting every possible IG handle from Webflow

We try to get every handle by realizing that if handle is blank the data should be in name:

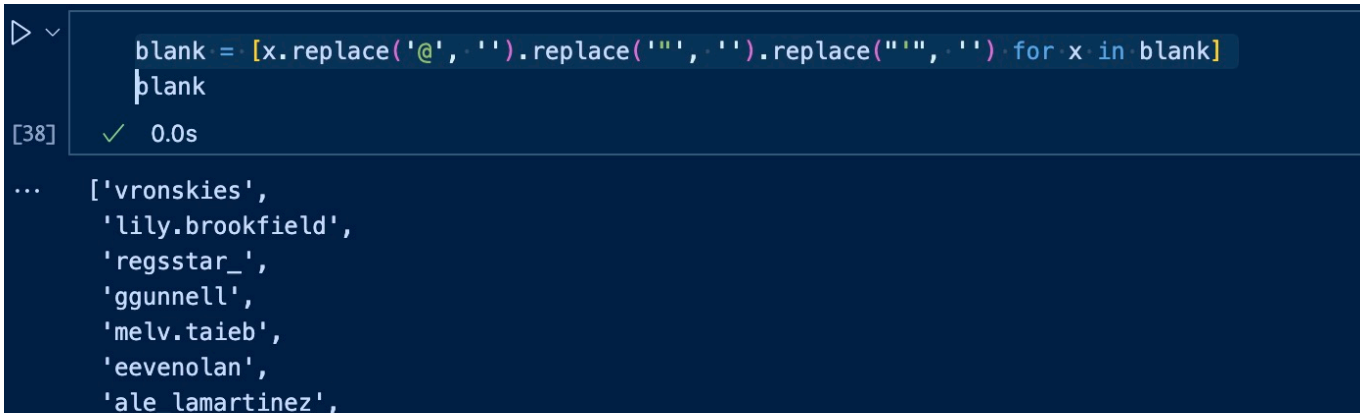
```
blank = []

for index, row in fuzzy.iterrows():
    if pd.isna(row['IG Handle']) or row['IG Handle'] == '':
        blank.append(row['Name'])
    else:
        blank.append(row['IG Handle'])

print(blank)
print(len(blank))
```



```
['vronskies', "'@lil
1019
```

Same number of rows as above, no data loss

Clean this output:



```
blank = [x.replace('@', '').replace('"', '').replace("'", '') for x in blank]
blank
```

```
[38]  ✓  0.0s
```

```
['vronskies',
 'lily.brookfield',
 'regsstar_',
 'ggunnell',
 'melv.taieb',
 'eevenolan',
 'ale lamartinez',
```

Best guess

```
import pandas as pd
from rapidfuzz import fuzz

def find_mismatched_handles(blank_list, df, threshold):

    # Extract the unique handles from dataframe (and remove any NaNs)
    df_handles = df["IG Handle"].dropna().unique().tolist()

    not_in_df = []
    for handle in blank_list:
        # Compute the best (maximum) ratio for 'handle' against all known df_handles
        best_ratio = max(
            fuzz.ratio(handle, h) for h in df_handles
        ) if df_handles else 0  # handle empty df_handles gracefully

        # If it's below the threshold, we consider it "not in the dataframe"
        if best_ratio < threshold:
            not_in_df.append(handle)

    return not_in_df


df = pd.read_csv('final.csv')
print(df.shape)
threshold = 90
mismatched = find_mismatched_handles(blank, df, threshold=threshold)
print("Handles not in dataframe (below similarity threshold):", mismatched)
print(len(mismatched))
```

Use this one:

(962, 13)
Handles not in dataframe (below similarity threshold of 90): ['d.g.006', 'Amira Hassan', 'alejandrafernandezmora', '', 'thanthar__', 'evelynmira99', 'samascosta', 'wedalmatar', '063maggie', 'kisokisoo', 'noemicworld', 'Deleted', 'test', 'nikahit', 'alinamariesophiew', 'arthurcrovetto', 'siegkrysthal', 'Hannah Kim', 'Liamfitandfun', 'Gems.haney', 'gabkropfl', 'ashleygallowayyy', 'yaseminnymanme.com', 'mara.gindorf', 'Seanpfc', 'mae.krn', 'luci.didonna', 'Ksksksk', 'Johnathan.Puls', 'hannahaltgassen', 'thtamphm', 'francoismartin', 'Juliampaz', 'zoefroxilia', 'martinclrmnt', 'Zalmeeb', 'andreadivalentina', 'cantthinkofausername000', 'danigroman', 'melissa.garay', 'Alicia_torriani', 'Ferr']
42

(962, 13) Handles not in dataframe (below similarity threshold of 95): ['regsstar', 'ggunnell', 'elinaktm', '_zzohaa', 'abbieg.', 'lilli12o1', '_.bex', 'alinxbr', '0ce1an', 'ananyyax', 'margotlds', 'saieemole', 'alinajmnz', 'ponitaty', 'ebaide.a', 'smitsvs', 'aliceadjr', 'farahq24', '75250frc', 'ferrenyw', 'elineckat', '_esidore', 'tntbkl', 'evaberk', 'i.g.m.h', 'prxshita', 'linsfikri', 'd.g.006', 'rositadhvr', 'aann.elle', 'susannaxh', 'vrqnika', 'itsekans', 'hanaabel', 'imaanes', 'nehrayaa', 'tara.va', 'hafsah.v', 'hira.nvh', 'roxyotero', 'xxloree', 'Amira Hassan', 'alejandra_fernandezmora', 'resu14700', 'laii.ia', '_luxylu', 'anatujo', 'cloe_stph', 'c.rsnne', 'mikdlvs', 'ypkva', 'lu.vdr', 'xufana', 'junetroan', 'alisandry', 'muyion', 'cindywxng', 'clem.hlg', 'maaaanina', 'alexuslg', '', 'saskiahkn', '_drialvez', 'annesowsr', 'begumoznl', 'yravasio', 'ygp0emi3', 'rae.h.kim', 'c.rsnne', 'samaabada', 'lmmdlw', 'thanthar_', 'marchetoo', 'menaaraim', 'Larisuh', 'mariapbv', 'tanaraflm', 'mialuh', 'emclean', 'dreasf', 'tn.aiya', 'evelynmira99', 'michi.zee', 'karla_b27', 'rigzum', 'samascosta', 'ttaniamb', 'megangoh', 'izzydut', 'snanyarko', 'carestonn', 'eltsks', 'firasdmk', 'sav.mello', 'wedalmatar', 'gdrouant', 'darcywt', 'pro.ffs', 'nourysul', 'fatma.lft', 'efrenaye', '063maggie',

```
'kisokisoo', 'mimz2005', 'mayarodic', 'kyleem4yy', 'gaiacrc', 'i_am_gabs', 'toniiwi', 'valiiib',
'brenaef', 'noemicworld', 'Deleted', 'annaorwin', 'emlsncl', '1vnaas', 'marie_lsc', 'lifeby.ck',
'na.d1ne', 'alinakwie', 'jaz_rihal', 'camila_ht', 'rhen_anne', 'boyanabh', 'catmgonz', 'moiragen',
'ritaoulds', 'panizjay', 'wildele', 'leaelil', 'idamcl', 'elvie_may', 'emmat', 'kyra.mvw', 'Hy0xae',
'chlfr__', 'nico.2108', 'dejjys', 'kyleem4yy', 'test', 'emilie_v5', 'coupdcut', 'mvdw.44', 'hadooni',
'Doris6522', 'thngoc.94', 'yeennhidg', 'shotr.an', 'rubydefa', 'timobenzz', 'amauryynf', 'voyushdv',
'alaiin', 'delphsr', 'maxjrnt', 'stevem_17', 'nikahit', 'alinamariesophiew', 'hugopkahn',
'arthurcrovetto', 'siegkrysthal', 'Hannah Kim', 'Liamfitandfun', 'Gems.haney', 'gabkropfl',
'ashleygallowayyy', 'yaseminnymanme.com', 'a.adwn', 'j4azminnn', 'mara.gindorf', 'noahzri',
'Seanpfc', 'aley.kd', 'mae.krn', 'luci.didonna', 'Aoibhin.gre', 'Ksksksk', 'Johnathan.Puls',
'lclaudiaa', 'daphnedlg', 'patrickbr', 'omarbarg', 'hannahaltgassen', 'thtamphm',
'francoismartin__', 'Juliampaz', 'zoefroxilia', 'martinclrmnt', 'noa_ibghi', 'enzofunke', 'Zalmeeb',
'diysly', 'coccadid', 'andreadivalentina', 'cantthinkofausername000', 'lvb812', 'danigroman',
'katvalee', 'melissa.garay', 'edobrof', 'Alicia_torriani', 'Ferr'] 202
```